remote sensing

MDPI

Article

Multi-Oriented Enhancement Branch and Context-Aware Module for Few-Shot Oriented Object Detection in Remote Sensing Images

Haozheng Su, Yanan You * and Sixu Liu

School of Artifical Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2017_liusixu@bupt.edu.cn (S.L.)
* Correspondence: youyanan@bupt.edu.cn

Abstract: For oriented object detection, the existing CNN-based methods typically rely on a substantial and diverse dataset, which can be expensive to acquire and demonstrate limited capacity for generalization when faced with new categories that lack annotated samples. In this case, we propose MOCA-Net, a few-shot oriented object detection method with a multi-oriented enhancement branch and context-aware module, utilizing a limited number of annotated samples from novel categories for training. Especially, our method generates multi-oriented and multi-scale positive samples and then inputs them into an RPN and the detection head as a multi-oriented enhancement branch for enhancing the classification and regression capabilities of the detector. And by utilizing the context-aware module, the detector can effectively extract contextual information surrounding the object and incorporate it into RoI features in an adaptive manner, thereby improving its classification capability. As far as we know, our method is the first to attempt this in this field, and comparative experiments conducted on the public remote sensing dataset DOTA for oriented object detection showed that our method is effective.

Keywords: few shot; oriented object detection; MOCA-Net; multi-oriented; context-aware

Citation: Su, H.; You, Y.; Liu, S. Multi-Oriented Enhancement Branch and Context-Aware Module for Few-Shot Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3544. https://doi.org/10.3390/rs15143544

Academic Editors: Xinghua Li, Fan Zhang, Bo Tang, Wei Yao, Zhongling Huang and Zongxu Pan

Received: 27 May 2023
Revised: 8 July 2023
Accepted: 11 July 2023
Published: 14 July 2023

1. Introduction

In the field of remote sensing image analysis and processing, object detection is a very important research interest. The emergence of deep convolutional neural networks (CNNs) has given rise to lots of remarkable CNN-based methods for object detection. These approaches can be classified into two main categories: horizontal object detection [1–4] and oriented object detection [5–8], based on the difference in bounding box representation. But these methods usually depend on using a large-scale and diverse dataset, which entails a laborious and time-consuming process to accumulate huge amounts of data. Thus, the demand to develop robust detection capabilities for new categories when provided with only a few annotated samples is becoming increasingly urgent in object detection.

Over the last few years, few-shot learning (FSL) has gained prominence as a novel research focus, with the goal of training models using few annotated samples. FSL methods mainly focus on the classification problem, such as the Siamese neural network [9], the matching network [10], and the prototypical network [11]. Based on these FSL methods, to require the model to not only recognize the object classes but also localize the objects in the image, few-shot object detection (FSD) has emerged as a prominent area of study in computer vision research. Usually, in the case of FSD, the dataset includes both base classes, which have numerous annotated samples, and novel classes, which have only a few annotated samples. Then, even with only a few annotated samples of novel classes and the aid of base classes, FSD models can more effectively detect novel classes than general detection methods. There are two approaches to categorizing the existing FSD methods: meta-learning based methods [12–14] and transfer-learning based

methods [15–17]. Specifically, meta-learning aims to learn prior knowledge from a series of tasks for new tasks, while transfer-learning methods transfer the existing model learned from an original task to a new task through some measures. These FSD methods have shown significant progress on natural image datasets, such as PASCAL VOC and MS COCO.

However, these few-shot object detection methods all belong to horizontal object detection, and the orientation information of objects cannot be captured solely by using horizontal bounding boxes, which severely limits the use of these methods in practical applications, such as prow detection [18] and object change detection [19] in remote sensing images. And the multi-oriented and dense distribution characteristics of objects in remote sensing images often lead to situations where several objects are tightly grouped and encompassed by a single horizontal bounding box, typically resulting in misaligned bounding boxes and objects. Therefore, oriented object detection that can effectively solve the above problems has developed rapidly in recent years. But, it requires the detector to have higher classification and regression capabilities for multi-oriented and multi-scale objects. Specifically, multi-oriented refers to the situation where objects have different orientations in the images resulting from a single perspective with more consistent object features. And this is different from the situation known as multi-angular, where the images resulting from multiple perspectives contain information from different sides of the object, leading to significant differences in object features. Multi-scale refers to the significant difference in the size of objects of the same category in the images, which can be seen as the difference in the number of pixels contained in the object. In few-shot oriented object detection, because of the lack of annotated samples for model training, the challenge is more severe. To visualize this challenge, we take the ground track field class in the public oriented object detection remote sensing dataset DOTA [20] as an example, and randomly select 30 samples as the few-shot case, while all labeled samples from the original dataset are used as original cases. As depicted in Figure 1, the orientation and scale distribution of objects in the few-shot case (blue column) are considerably sparser than those in the original case (red column) of abundant training data. This challenge makes it more difficult for few-shot detectors to learn the ability to detect multi-oriented and multi-scale objects, and the issue of enriching the orientation and scale space using only a few annotated samples is still an open question in need of a solution.

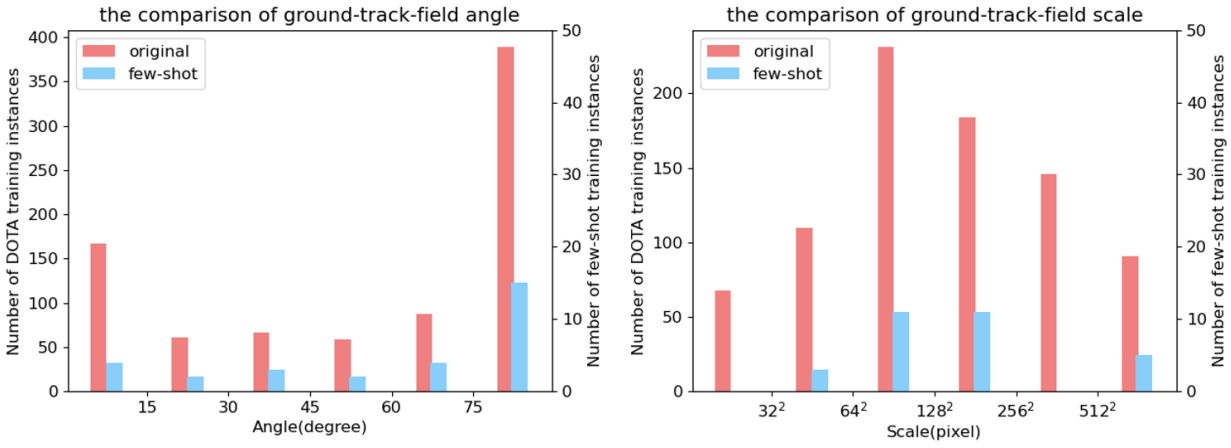

Figure 1. Comparison of orientation and scale distribution in DOTA (original) and a 30-shot subset of DOTA (few-shot). The ground track field class is used as an example.

Moreover, oriented object detection methods [6,21] often utilize rotate RoI alignment (RRoI align) to extract oriented proposals from the feature map for subsequent classification, which mainly contains the foreground information of the object, while there is little contextual information about the object and its surrounding environment. Researchers generally believe that the contextual information around the object is of little significance

to oriented object detection [6,22] when an abundance of annotated samples are available. But, in few-shot oriented object detection, it is not enough to train the detector to effectively distinguish each object depending only on the foreground information. So, how to make use of other information in the few annotated samples is a very important problem. Most types of objects have specific backgrounds that contain different contextual information; as shown in Figure 2a,b, the surrounding environment between ships and vehicles is different, and this clue can be used to help classification [23]. However, as depicted in Figure 2c,d, because surrounding environments are changeable and objects are often densely distributed in remote sensing images, it is not effective to simply expand the bounding box to add contextual information [6]. Thus, effectively extracting and integrating the contextual information with the foreground information remains a significant challenge.

To tackle these challenges, we introduce a few-shot oriented object detection method: MOCA-Net. Specifically, based on the latest oriented object detection method, Oriented R-CNN [22], our method introduces multi-oriented enhancement branch and context-aware modules. To solve the problem of orientation and scale distribution sparsity, inspired by MPSR [24], we propose a multi-oriented enhancement branch, adding a new branch to generate multi-oriented and multi-scale positive samples to assist training. During model training, this enhancement branch not only classifies the generated multi-oriented and multi-scale objects but also performs oriented bounding box regression in a region proposal network (RPN) and the detector head, which can enhance both the classification and the regression capabilities of the detector for objects with different orientations and scales. This branch shares parameters with the basic detector, and we actively select feature pyramid network (FPN) [25] stages and proposals to avoid introducing a large number of improper negative samples. Meanwhile, we also propose a context-aware module for object classification that introduces contextual information between the object and the surrounding environment. The module expands the areas of proposals to include the background as context-aware areas, obtains their features through RRoI align, and then determines the relevance between each context-aware feature and its corresponding RoI features. Then, the contextual information is adaptively integrated into the RoI features based on their relevance. And the module makes full use of the contextual information in the external horizontal bounding box, selects the top/bottom/left/right half part as the contextual subregions, which contain different contextual information about different parts of the object and the background, then concatenates and fuses these corresponding features after RRoI align, and finally integrates them into RoI features. We tested our method on the public oriented object detection dataset DOTA [20], and general oriented object detection methods, such as RoI transformer [6] and ReDet [8], and existing FSD methods, such as Meta R-CNN [13] and TFA [16], were adopted to compare with our method and proved its effectiveness. Our key contributions, summarized as four main points, can be outlined as follows:

1. We introduce a few-shot oriented object detection method with a multi-oriented enhancement branch and context-aware module. As far as we know, we are the first to introduce few-shot object detection research into the field of oriented object detection.
2. To deal with this problem of oriented and scale distribution sparsity, we propose a multi-oriented enhancement branch to enrich the orientation and scale space of the object.
3. We propose a context-aware module that introduces the context mechanism into few-shot oriented object detection.
4. This is the first time that we have built three different base/novel settings specifically for few-shot oriented object detection on the large-scale remote sensing dataset DOTA. Furthermore, we present baseline results that can serve as a benchmark for future studies.

(a) a vehicle on the road (b) a ship on the sea (c) many vehicles in the parking lot (d) many ships moored at the port

Figure 2. Ship class and vehicle class in different environments, which contain different contextual information.

The subsequent portions of this article are organized in the manner described below. Section 2 offers a brief overview of the relevant literature regarding our method. Section 3 provides a comprehensive description of our proposed approach, MOCA-Net. In Section 4, we present the experimental setup and comprehensive results. We further discuss our method in Section 5. In the end, we summarize the article in Section 6.

2. Related Work

2.1. Oriented Object Detection

Considering the lack of orientation information in horizontal object detection, there has been a growing focus on oriented object detection recently, resulting in the emergence of many outstanding methods based on classical horizontal object detection frameworks [1–4]. The oriented bounding box containing orientation information is used to represent multi-oriented objects in oriented object detection. Depending on how the oriented bounding box is represented, two different approaches can be used to categorize oriented object detection methods: the detection methods based on orientation representation [5–8,21,22,26] and the detection methods based on vertex representation [19,27,28].

Orientation representation uses the rotated angle (θ) of the bounding box and, together with the midpoint (x, y), length, and width (w, h), it forms a representation of the oriented bounding box (x, y, w, h, θ). Based on this representation, RRPN [21] and R2PN [26] use the rotating RPN network to introduce rotating anchors of different orientations, sizes, and aspect ratios into object detection. However, a large number of anchors brings about computation complexity and greater memory consumption. To decrease the quantity of anchors, a RoI transformer [6] obtains the rotating RoIs from the horizontal RoIs obtained by RPN. In view of the slow detection speed of multi-stage detectors, Han et al. [7] proposed Sa2net based on the single-stage detection algorithm RetinaNet, which improves the detection speed while maintaining excellent accuracy. In order to eliminate the influence of object orientation change on feature extraction, ReDet [8] uses rotation invariant backbone and rotation invariant ROI align to obtain the rotation invariant feature of the object. The latest method, Oriented R-CNN [22], uses Oriented RPN to directly generate the rotation suggestion box. Compared with a RoI transformer, it effectively reduces the computation complexity.

Vertex representation indicates the oriented object by marking four vertices of the quadrilateral $(x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_4)$. Compared with the orientation representation, it can represent any quadrilateral. Liao et al. [27] proposed an irregular long convolution to detect high aspect ratio objects. Gliding Vertex [28] adds the offset values of four vertices and judges whether they are oriented in the bounding box regression output of Faster R-CNN. FR EST [19] generates 16 key point thermal maps on the horizontal RoIs, where every four adjacent pixels represent a corner.

But these studies all depend on large-scale and diverse oriented object detection datasets for training, such as DOTA [20]. Obtaining a large number of annotated samples for novel classes often requires substantial financial resources. And limited annotated

samples during the training process of a detector can contribute to the occurrence of overfitting, causing a sharp decline in its generalization capability [12].

2.2. Few-Shot Object Detection

Because the existing object detection methods often need a large-scale dataset, few-shot object detection (FSD), which only needs to provide a few annotated samples of novel classes for training, has attracted more and more attention, and many well-designed methods have been proposed. Two groups can be distinguished among the existing FSD methods: meta-learning-based methods [12–14,29–31] and transfer-learning-based methods [15–17,24,32], representing different approaches to address the few-shot detection challenge.

Specifically, meta-learning, often referred to as learning to learn, aims to learn prior knowledge from a series of tasks for the model training of new tasks with only a few annotated samples. FSRW [12] and Meta R-CNN [13] extract the attention vector of each class from support images to weight the feature maps. RepMet [14] combines a detection backbone network, embedded space, and object classification, then classifies by calculating the distance between the embedded space vector and representative vectors of each class. FSOD [29] matches objects in query images by using support images of each class. It does not need to train the model again when predicting the new class. To resolve the limited generalization ability of RPN, Hu et al. [30] proposed a coarse prototype matching network to replace the traditional linear classifier with a nonlinear classifier based on metric learning. FCT [31] introduces a transformer into few-shot object detection and conducts feature interaction between query images and support images at each stage.

In contrast, transfer-learning-based methods enable the transfer of learned knowledge from an original task to a new task by utilizing some measures, and only retraining the existing model can achieve an effect not inferior to that based on meta-learning. LSTD [15] proposed transfer knowledge regulation and background depression regulation to assist model transfer. TFA [16] achieved excellent results by transferring only the last layer of the Faster R-CNN detector. FSCE [17] adds a contrastive learning branch on the basis of TFA to strengthen RoI, and the comparative branch measures the similarity between the proposed encoding. SRR-FSD [32] combines the semantic relationship and visual information between each class and introduces explicit relation reasoning into the learning of few-shot object detection. MPSR [24] proposed multi-scale positive sample refinement to deal with object-scale distribution sparsity. Due to the simplicity, efficiency, and excellent performance of transfer-learning, our method focuses on this approach.

However, these methods are all horizontal object detection, which cannot provide orientation information. Due to the limitations of horizontal object detection in fields such as prow detection and the advantages of oriented object detection in remote sensing images, the significance of few-shot oriented object detection becomes apparent in addressing these challenges, but there is still a lack of method in the field. Moreover, due to arbitrary orientations and various scales of oriented bounding boxes, few-shot oriented object detection becomes more challenging when utilizing the oriented bounding box as opposed to the horizontal bounding box.

2.3. Few-Shot Learning in Remote Sensing Images

In light of the crucial applications of remote sensing images, researchers are devoting their efforts to exploring few-shot learning algorithms customized for this context, especially in specific remote sensing fields such as synthetic aperture radar automatic target recognition (SAR-ATR) and ship detection. Cai et al. [33] proposed an improved prototypical network (PN) based on spatial transformation named ST-PN, which is applied to SAR-ATR. M Rostami et al. [34] introduced a new algorithm to solve the problem of ship classification with limited data. Tai et al. [35] proposed a novel few-shot transfer learning method with a connection-free attention module for SAR image classification. Wang et al. [36] also proposed a new few-shot SAR ATR method based on Conv-BiLSTM Prototypical Networks. Ai et al. [37] proposed a SAR image target detection method based

on a multi-level depth learning network that fuses the high-level target depth feature. Zhang et al. [38] proposed a few-shot ship detection algorithm based on the YOLO algorithm, which achieved a better result. Zhang et al. [39] introduced a remote sensing few-shot object detection method based on text semantic fusion relation graph reasoning (TSF-RGR), which learns various types of relationships from common sense knowledge. Chen et al. [40] proposed a novel few-shot SAR object detection framework, which is built upon the meta-learning architecture, introducing few-shot object detection into SAR images for the first time.

In short, compared to the methods described above, our approach is committed to solving the task of few-shot oriented object detection in remote sensing images and is the first attempt in this field.

3. Method

3.1. Problem Definition

In this study, we adopt the commonly used few-shot detection settings [12,13,16]. Suppose that we are provided with two datasets: $D_{base} = \{(x_{base}, y_{base})\}$, including base classes C_{base} and $D_{novel} = \{(x_{novel}, y_{novel})\}$, and including novel classes C_{novel}, where $C_{base} \cap C_{novel} = \emptyset$, x represents the image in the dataset, and y represents the corresponding label. D_{base} contains a large amount of annotated samples, while D_{novel} is usually set as K-shot, which means that each novel class only contains K (e.g., $K = 3, 5, 10$) annotation samples. The training process of the model typically involves a two-stage approach. In the first stage, only D_{base} is used to train the basic detector, and in the second stage, D_{base} and D_{novel} are used to train together. To avoid the problem of sample imbalance, each base class selects the same number of samples as each novel class. The final goal is to use base classes with lots of annotated samples to assist in the detection of novel classes with only a few annotated samples, resulting in a detector that can detect not only base classes but also novel classes.

3.2. Overview of the Proposed Method

Figure 3 presents the framework of our method, MOCA-Net. And the core of MOCA-Net is introducing a multi-oriented enhancement branch (MOE) and a context-aware module (CAM). Based on the two-stage detector Oriented R-CNN, our method first extracts each object in the input image and converts them to various orientations and scales, and then sends the generated images to RPN and detection head as the multi-oriented enhancement branch. To avoid introducing lots of improper negative samples, in RPN, different from the anchor matching of the basic detector, the corresponding scale level of the feature maps is manually selected, the matched positive samples are input into both the RPN classifier and RPN regression, and we add the loss of output results into the RPN loss to enhance the location and regression capabilities. In the detection head, we use the ground truth of the object with a random offset to extract the features by RRoI align, simultaneously input them into the object classification and bounding box regression, and then add the loss to the RoI loss. The multi-oriented enhancement branch and basic detector use shared parameters. Meanwhile, our method adds the context-aware module to enhance the classification capability in the detection head. To obtain the contextual information between the object and its background, we first expand the length and width of oriented proposals by r times (here, r is 1.5), use RRoI align to obtain context-aware features from feature maps, and build a relevance matrix between them and the corresponding RoI features; then, this matrix guides the integration of contextual information into the RoI features. And to obtain the contextual information in the corresponding external horizontal bounding box, we divide the rectangle into four local areas, concatenate their features after RRoI align and fuse them through 1×1 convolution, then combine them into RoI features. Through the utilization of the context-aware module, the classification difficulty is effectively reduced by mining contextual information from a limited number of annotation samples.

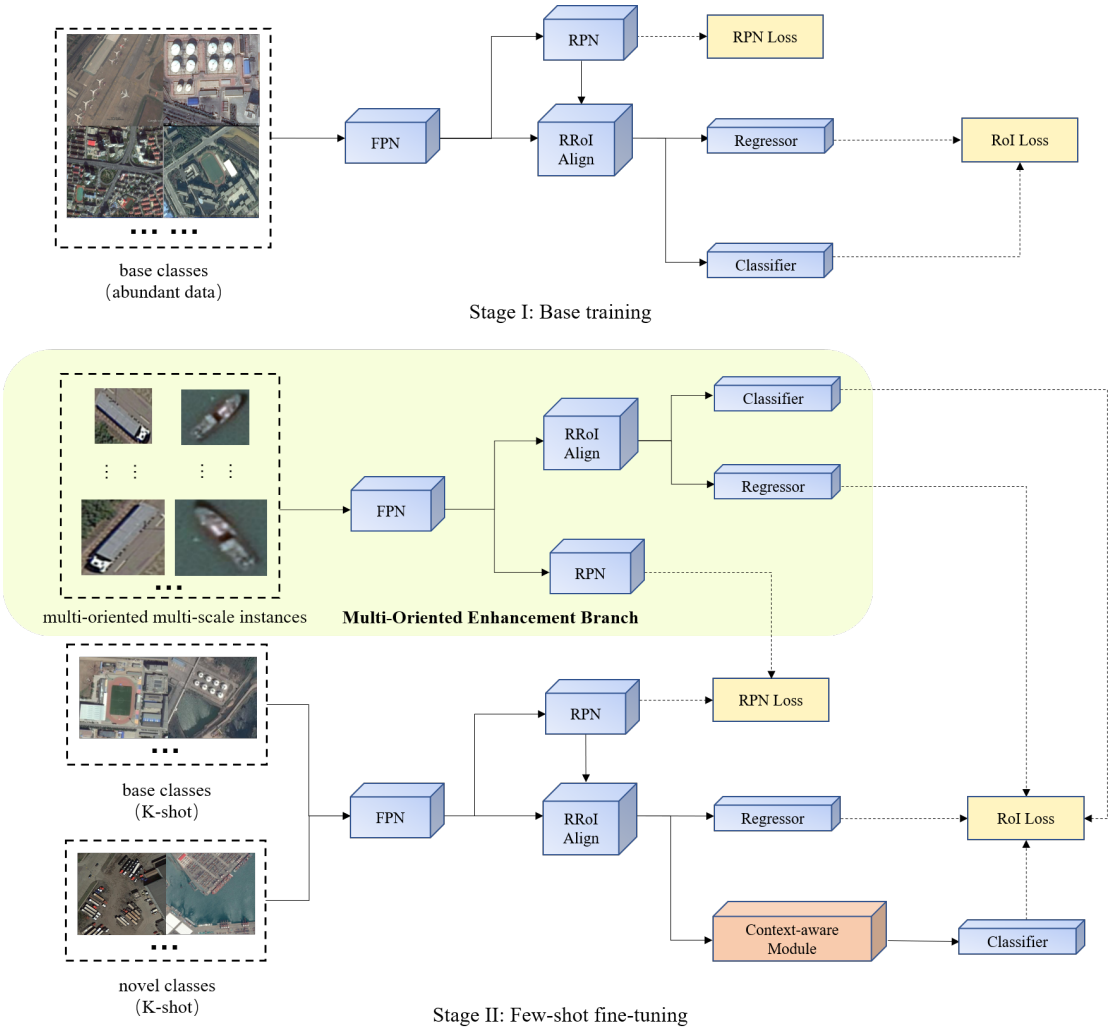

Figure 3. Framework of our proposed method, MOCA-Net. The core of our method is introducing the multi-oriented enhancement branch to enrich the orientation and scale space of the object and the context-aware module to obtain contextual information around the object. And our method adopts a two-stage training strategy: base training and few-shot fine-tuning.

3.3. Training Strategy

As shown in Figure 3, our approach follows a two-stage training strategy, similar to the common FSD method [16,24], which is composed of the first stage of base training and the second stage of few-shot fine-tuning. During the first stage of base training, for training the basic detector, we only use the D_{base} with abundant annotation samples, and in this stage, neither the multi-oriented enhancement branch nor the context-aware module are utilized. In the second stage of few-shot fine-tuning, by the transfer of prior knowledge learned from the base categories in the basic detector, our approach can detect objects of the novel classes with few annotated samples for training, and our model is trained jointly using data from the base classes and the novel classes to achieve concurrent detection of all classes $C_{base} \cup C_{novel}$. Because the number of annotated samples per novel class is limited to K, we randomly choose K annotated samples per base class from D_{base}, ensuring an equitable representation of samples for both the novel and base classes. Moreover, for each annotation sample, we use a square window to extract each object in the image with a minor random shift and transform its orientation and scale to obtain multi-oriented and multi-scale objects as the input of the multi-oriented enhancement branch. And during model training, we use the multi-oriented enhancement branch and context-aware module together to assist model transfer.

3.4. Multi-Oriented Enhancement Branch

To enrich the scale and orientation space of the object, data enhancement methods such as transforming image size and image orientation [1,3,6,22] have been widely used in object detection with a large number of annotated samples. However, these strategies will increase the number of improper negative samples, which will seriously affect the training of the model when positive samples are scarce, resulting in poor effects [24].

To overcome this challenge, shown in Figure 4, we introduce a multi-oriented enhancement branch aimed at enriching the orientation and scale space of objects without significantly increasing the number of improper negative samples. Firstly, we use a square window (the side length is the same as the longest side of the object's external horizontal bounding box) to extract each object in the image with a minor random shift, resize its scale to $\{32^2, 64^2, 128^2, 256^2, 512^2\}$, resembling the anchor design of the basic detector, and the orientation of the object is rotated randomly for each scale. Then, all the generated images are sent to the RPN and the detection head as the multi-oriented enhancement branch to assist in the training of the model.

In the RPN, the anchor scales corresponding to the multi-level feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ of the FPN are $\{32^2, 64^2, 128^2, 256^2, 512^2\}$, respectively. Since each image contains only one instance, this is different from the standard detection situation, where each image usually contains multiple objects. Therefore, applying anchor matching directly on cropped single objects can be inefficient and lead to an accumulation of improper negative samples, which can have detrimental effects on the detector's performance. In contrast to the conventional anchor matching approach, Figure 4 shows our manual selection process, where the scale level of the feature maps is chosen to match the scale of the input image. Because the RPN in Oriented R-CNN not only locates the object but also completes the regression of the horizontal anchor to the oriented proposal, and the regression task is related to the subsequent RRoI align, the regression task is as important as the classification task. So, unlike MPSR, we not only use the matched positive samples for RPN classifiers but also for RPN regression. The output of the RPN networks in the multi-oriented enhancement branch includes classifier scores and oriented proposals' predictions. The oriented RPN loss function of our method is as follows:

$$L_{RPN} = \frac{1}{M_{RPN} + N_{RPN}} \sum_{i}^{M_{RPN}+N_{RPN}} L_{cls}^{i} + \frac{1}{M_{RPN} + N_{RPN}} \sum_{i}^{M_{RPN}+N_{RPN}} L_{Preg}^{i} \tag{1}$$

where M_{RPN} denotes the count of positive anchor samples that are chosen for enhancement, and N_{RPN} represents the total number of selected anchors in the basic detector. For the ith anchor in a mini-batch, L_{cls}^{i} denotes the binary cross-entropy loss, differentiating between background and foreground classes, and L_{Preg}^{i} is the smooth L_1 loss defined in [22].

In the detection head, we directly use the ground truth of objects to extract RoI features from feature maps by RRoI align. As with the operations in RPN, we not only use the RoI features for object classification but also for bounding box regression. Meanwhile, we found that random clipping with a minor offset of the ground truth can effectively improve the generalization of classification and regression. We define the loss function for our detection head as:

$$L_{RoI} = \frac{1}{N_{RoI}} \sum_{i}^{N} L_{Kcls}^{i} + \frac{1}{N_{RoI}} \sum_{i}^{N} L_{Rreg}^{i} + \lambda \left(\frac{1}{M_{RoI}} \sum_{i}^{M} L_{Kcls}^{i} + \frac{1}{M_{RoI}} \sum_{i}^{M} L_{Rreg}^{i} \right) \tag{2}$$

where M_{RoI} denotes the count of chosen RoIs in the multi-oriented enhancement branch, and N_{RoI} represents the total number of RoIs in a mini-batch. For the ith RoI in a mini-batch, L_{Kcls}^{i} refers to the log loss computed over K classes, and L_{Rreg}^{i} represents the smooth L_1 loss.

The multi-oriented enhancement branch and basic detector use shared parameters for parallel training. Therefore, the multi-oriented enhancement branch will not add additional parameters during training. Meanwhile, the multi-oriented enhancement branch

is only used in the training stage. Because after the model training the multi-oriented enhancement branch will be removed, this branch will not incur additional time costs in the detection stage.

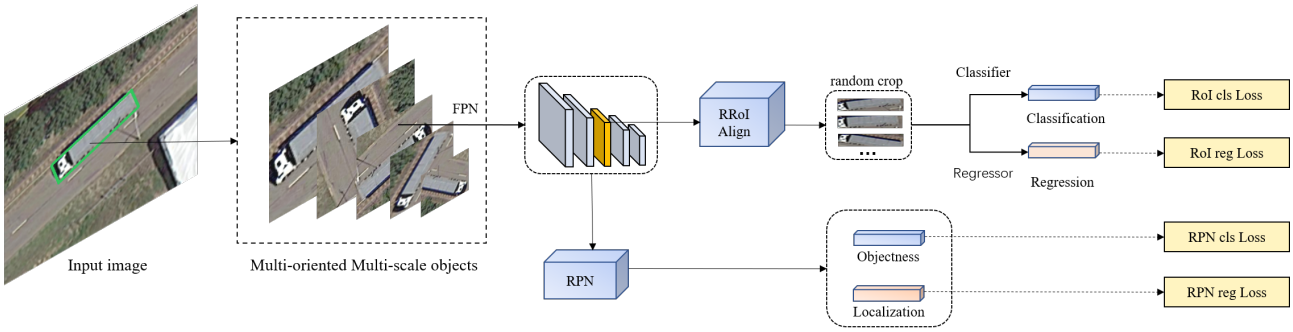

Figure 4. Details of the multi-oriented enhancement branch. For an input image, objects are extracted and then resized to varying orientations and scales within the multi-oriented enhancement branch. By feeding the generated images into the FPN, we identify and utilize specific features to enhance the classification and regression capabilities of both the RPN and the detection head.

3.5. Context-Aware Module

Since the oriented proposals are close to the object edge, the RoIs predominantly include the foreground of the object, leading to a lack in the capture of the contextual information surrounding it, which may contain useful information for classification. This contextual information is particularly important when the labeled samples are scarce. To address the issue of classification challenges, we present the context-aware module, which is incorporated into the detection head. As shown in Figure 5, this module is divided into contextual information extraction in the background and contextual information extraction in the horizontal bounding box. By extracting and integrating the contextual information in the background and the external horizontal bounding box, classification confusion can be reduced.

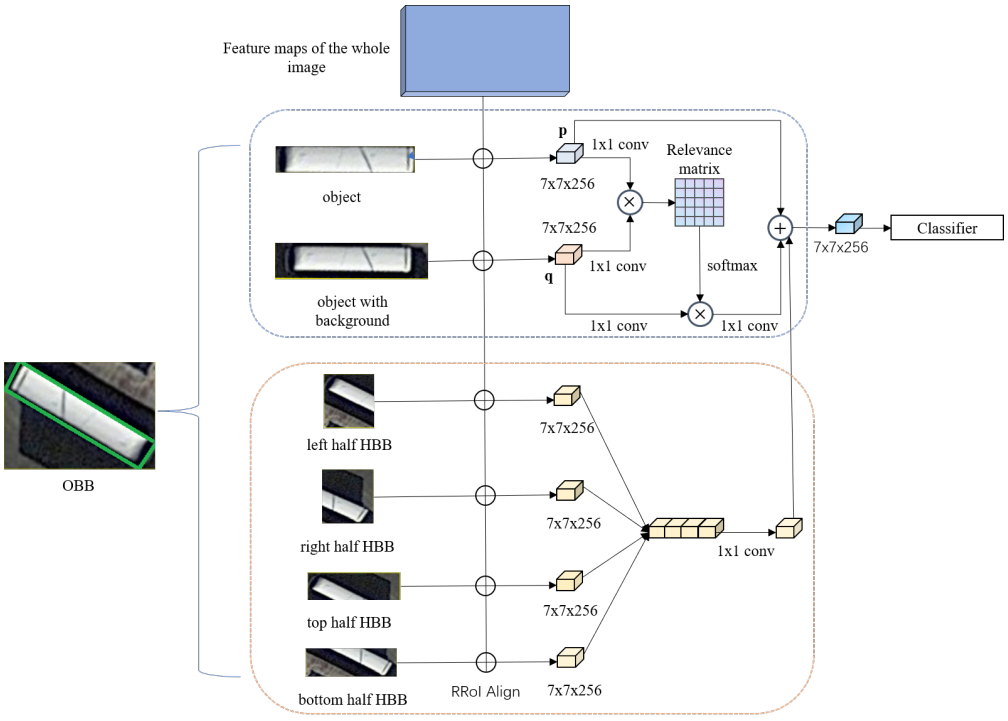

Figure 5. Details of the context-aware module. The module is divided into contextual information extraction in background (in the blue box) and contextual information extraction in horizontal bounding box (in the orange box).

Contextual information extraction in the background. Referring to the method we previously proposed in [41], we have proven that the self-attention mechanism can effectively extract contextual information from the background around the object to improve detection accuracy. But, the previous method we proposed was applied to few-shot horizontal object detection, and here we migrated its core module to few-shot oriented object detection for obtaining the contextual information between the oriented object and its background. We briefly describe the processes of the core module as follows:

First, we expand the length and width of the oriented proposals by r times (r is 1.5 here) to include the background around the object, and then extract the context-aware features $\boldsymbol{p} \in R^{N\times C\times H\times W}$ and the corresponding RoI features $\boldsymbol{q} \in R^{N\times C\times H\times W}$ through RRoI align, where N represents the mini-batch size, C denotes the dimension of the features, and H and W represent the height and width of the feature maps.

Then, we adopt the commonly used dot product kernel to establish the relationship between the context-aware features and RoI features. Through this process, a relevance matrix $F(\boldsymbol{p}, \boldsymbol{q})$ is created, which is formulated as follow:

$$F(\boldsymbol{p}, \boldsymbol{q}) = \varphi(\boldsymbol{p})^T \times \varphi^{'}(\boldsymbol{q}) \tag{3}$$

where φ and $\varphi^{'}$ represent two different 1×1 convolutions, which realize dynamic learning of similarity through a linear transformation of features. After obtaining the relevance matrix, we use softmax normalization to obtain the weighted matrix W_{ij} for obtaining the importance of each contextual information,

$$W_{ij} = \frac{exp(F(\boldsymbol{p}, \boldsymbol{q}))}{\sum_j exp(F(\boldsymbol{p}, \boldsymbol{q}))} \tag{4}$$

then use the weighted matrix W_{ij} Weighting $\boldsymbol{p}$ to obtain the weighted features V_{BG} of the background contextual information,

$$V_{BG} = \Phi\left(W_{ij}\Phi^{'}(\boldsymbol{p})\right) \tag{5}$$

where Φ and $\Phi^{'}$ represent two different 1×1 convolutions to increase learning flexibility.

Contextual information extraction in external horizontal bounding box. In general oriented object detection, the horizontal bounding box serves as the basis for generating the corresponding oriented proposals in oriented object detection, and is not considered for further classification and regression processes in the detection head. However, the horizontal bounding box contains not only the foreground of the object but also its surrounding environment. In addition, the utilization of multi-region feature fusion for the object enhances the feature robustness, enabling more precise classifications as it incorporates more localized details of the object and its surrounding environments [42]. Therefore, four local regions, namely the top, bottom, left, and right half parts, are created by dividing the rectangle. Each region only retains half of the horizontal bounding box. Rather than extracting features directly from the image, the features of these regions are derived from the corresponding regions on the feature maps generated from the whole image. Following that, RRoI align is used to change these features to a consistent size. Subsequently, these features are concatenated and fused through a 1×1 convolution, which facilitates the integration of features from diverse channels, resulting in improved nonlinearity without compromising resolution [43]. In the end, the contextual information features, V_{HBB}, with the same dimensions as RoI features, are obtained.

After obtaining the background context features V_{BG} and the horizontal bounding box context features V_{HBB}, these features are added into RoI features together for subsequent classification. Different from the classification related to the category, because the regression task in the detection head should make the bounding box contain only the foreground as much as possible, adding the information of the surrounding environments will interfere

with the subsequent regression. Therefore, the regression settings in the basic detector are retained without using the context-aware module.

4. Experiments and Results

4.1. Datasets and Setting

In this paper, we adopt a widely used large-scale remote sensing dataset, DOTA [20], to test the effect of our proposed method. The DOTA dataset consists of 2806 optical aerial images, and these images, with multiple resolutions, are obtained from various sensors and platforms, such as Google Earth. It contains 188,282 objects with oriented bounding boxes (task 1) and horizontal bounding boxes (task 2), covered by the following 15 object classes: Bridge (BR), Harbor (HA), Ship (SH), Plane (PL), Helicopter (HC), Small vehicle (SV), Large vehicle (LV), Baseball diamond (BD), Ground track field (GTF), Tennis court (TC), Basketball court (BC), Soccer-ball field (SBF), Roundabout (RA), Swimming pool (SP), and Storage tank (ST), which exhibit significant variations in scale, orientation, and aspect ratio. In this dataset, the image size is large: from 800×800 to 4000×4000, so we cut the original image to 1024×1024 patches. The clipping step is set to 824, which indicates that two patches have an overlap of 200 pixels. This dataset is split into three distinct sets: the training set, the validation set, and the testing set. The evaluation of detection accuracy requires the submission of detection results on the test set to the DOTA evaluation server, and there is a limit on the number of submissions in one day. However, the FSOD experiment generally uses N-split and K-shot settings, which will produce $N \times K$ detection results at once. The evaluation server has difficulty meeting the test requirements. Therefore, unlike the common oriented object detection methods [6,22], which use a training set and a validation set for training and a testing set for testing, we use a training set for training and a validation set for testing.

To establish the experiment setting for few-shot object detection, we constructed three base/novel class splits, in which the novel classes do not overlap with each other. For each split, we randomly selected five classes in the dataset as the novel classes and the remaining ten classes as the base classes. Table 1 shows the detailed division. For training, we selected $K = 3, 5, 10, 20, 30$ objects for the novel class, which are randomly selected from the training set. We test our methods on the validation sets that contain both base classes and new classes.

To evaluate our method, the evaluation metric we selected is mean average precision (mAP), a commonly used method for assessing general object detection. The performance evaluation is conducted using the PASCAL VOC2007 development kit on the mmrotate platform [44]. We conducted three repetitions of each experiment, aiming to achieve stable outcomes.

Table 1. Three different base/novel class split settings in our experiments.

Split	Novel Classes					Base Classes
1	Plane (PL)	Large vehicle (LV)	Ship (SH)	Ground track field (GTF)	Harbor (HA)	rest
2	Storage tank (ST)	Baseball diamond (BD)	Basketball court (BC)	Tennis court (TC)	Roundabout (RA)	rest
3	Bridge (BR)	Small vehicle (SV)	Helicopter (HC)	Soccer-ball field (SBF)	Swimming pool (SP)	rest

4.2. Implements Details

We built our model and generated our experimental results on the mmrotate platform [44]. Because our proposed method is based on Oriented R-CNN, we also used the same ResNet50 [45] pre-trained model on ImageNet as the backbone network, and our hyperparameters use their initial settings. The only change implemented is the adoption of the class-agnostic regression task in the detection head. An RTX 3090 is used for training and testing our model. During training, we used an SGD optimizer with a momentum coefficient of 0.9 and a weight attenuation of 0.0001. In the first stage of base training, we used a batch size of four to train our model. The model was trained for 20 k, 6 k, and 4 k iterations, and the learning rates were 0.005, 0.0005, and 0.00005, respectively. In the second

stage of few-shot fine-tuning, the replacement was limited to the last fully connected layer used for classification in the detection head. The new fully connected layer was initialized with random values, and all layers remained unfrozen throughout the fine-tuning process. Considering the small number of annotated samples in the second stage, to mitigate the risk of overfitting, the training times and learning rate in this stage were appropriately reduced. The model was trained for 2000, 600, and 400 iterations, and the learning rates were 0.001, 0.0001, and 0.00001, respectively.

4.3. Comparing Methods

To demonstrate the effectiveness of our method, we conducted a comparative experiment against a variety of general oriented object detection methods, including FR-O [20], Giding Vertex [28], R3Det [46], Sa2net [7], RoI Transformer [6], ReDet [8], and Oriented R-CNN [22], which is the basic detector of our method. Like our method, these approaches follow a two-stage training strategy: in the first stage, we train the detection model using a large dataset that only consists of base classes; in the second stage, the model is fine-tuned by the dataset combining both base classes and novel classes.

Few-shot object detection has developed rapidly in recent years, and a lot of excellent methods have emerged and achieved excellent results. However, these methods are based on horizontal object detection datasets. To further substantiate the effectiveness of few-shot oriented object detection in remote sensing images, we selected several FSD methods for comparison, including meta-learning based methods, Meta R-CNN [25], FsDet [47], and AttentionRPN [23], and transfer-learning based methods, TFA [16], MPSR [24], and FSCE [17]. Although these horizontal few-shot object detection methods are not directly compatible with the oriented bounding box data in DOTA (task 1), they can be used for the horizontal bounding box data in DOTA (task 2). The horizontal bounding box of task 2 is the external rectangle corresponding to the oriented bounding box of task 1, and the detection metric is also mAP, similar to ours.

4.4. Experimental Results and Comparisons

The detection performance of our method and the general oriented object detection methods on different splits and shots in the DOTA dataset is reported in Table 2. The results clearly indicate a significant decrease in detection accuracy for general oriented object detectors when there are only a few annotated samples. This situation clearly indicates that, when the object features provided to the model during training are too few, the model will encounter overfitting problems and its generalization ability will also be significantly weakened. Our method, MOCA-Net, performs best in all splits and all shots and is significantly superior to the latest oriented object detection methods like ReDet and Oriented R-CNN. The mAP is improved by averages of 5.5%, 8.2%, and 4.1% in split1, split2, and split3 settings, respectively, compared to the baseline Oriented R-CNN, which demonstrates the effectiveness of our method.

Table 2. Detection results for the novel classes, and mAP (%) is used as an evaluation measure. Red and blue represent the best and the second best.

	Split 1					**Split 2**					**Split 3**				
Method/Shot	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**
FR-O	5.8	9.4	10.9	15.1	18.4	4.7	10.8	21.6	33.2	38.4	3.1	9.2	11.5	20.7	22.0
Giding Vertext	5.5	10.3	12.1	16.8	17.8	3.2	9.6	19.9	30.5	34.2	1.7	6.4	13.6	17.4	18.6
R3Det	4.8	10.1	10.5	14.6	16.0	6.9	12.2	14.9	26.9	31.7	2.9	9.5	11.8	17.9	21.1
Sa2net	2.3	7.7	10.2	15.4	18.8	3.3	10.9	17.1	29.8	35.8	3.4	7.9	14.3	18.9	19.5
RoI Transformer	6.2	10.4	12.8	18.4	**20.9**	7.4	13.6	21.8	33.9	40.6	3.7	**10.6**	14.5	**21.7**	**22.8**
ReDet	5.7	10.5	**14.8**	**18.9**	20.3	9.5	12.1	17.8	**34.4**	**42.1**	4.8	9.9	14.1	20.9	21.5
Oriented R-CNN	**7.8**	**12.1**	14.2	18.1	20.4	**10.1**	**14.2**	**22.8**	33.6	39.6	**5.6**	10.4	**15.3**	21.4	22.6
ours	**11.2**	**17.9**	**22.1**	**23.7**	**26.1**	**20.4**	**26.7**	**29.1**	**39.6**	**45.7**	**10.3**	**16.2**	**18.4**	**24.6**	**26.3**

The experimental results comparing our method with the FSD method are shown in Table 3. It can be seen that our method, MOCA-Net, performs much better than the other FSD methods in all splits and all shots, and compared to FSD methods, our approach increased mAP by more than 5%, demonstrating the advantage of our method in few-shot oriented object detection. For the FSD methods shown in Table 3, the performance is similar or even lower compared to the latest general oriented object detection methods such as Oriented R-CNN, and some FSD methods such as TFA and Meta R-CNN, which are effective in nature images but have poor detection performance on remote sensing images, with a mAP that is less than 20% in a 30-shot setting.

We counted the detection results with average precision (AP) of each novel class of each split under 30 shots to show the performance of our method in each class. The details are shown in Table 4. It is evident from the table that our method, MOCA-Net, achieved the best and the second best in most categories, and other methods only achieve the best in one or two categories, such as the Sa2net method in the small vehicle (SV) category. And, compared to the baseline Oriented R-CNN, our method increased mAP by an average of 5.2%, further demonstrating its effectiveness.

Table 3. Detection results for the novel classes on task 1 (oriented bounding box) and task 2 (horizontal bounding box) of DOTA and mAP (%) are used as evaluation measures. Red and blue represent the best and the second best.

		Split 1					Split 2					Split 3				
	Method/Shot	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**
Task 2	TFA	7.8	11.5	12.6	14.2	14.8	4.5	9.4	10.8	16.2	18.6	5.2	7.4	11.9	13.2	14.5
	Meta R-CNN	6.7	7.8	10.2	11.3	13.6	5.2	8.9	19.3	22.5	29.8	4.3	6.2	9.6	11.7	13.8
	FsDet	6.9	9.2	10.8	11.2	12.3	4.6	7.9	18.1	21.3	28.7	4.9	6.7	10.4	12.6	13.1
	AttentionRPN	5.4	7.6	9.8	10.4	11.2	3.8	7.2	16.4	20.4	25.2	3.1	5.8	9.3	12.2	13.4
	MPSR	8.8	12.6	13.1	20.4	22.7	4.3	12.5	19.7	31.6	37.3	4.1	11.7	13.2	19.3	20.8
	FSCE	7.9	11.8	15.1	19.2	22.5	10.7	14.8	19.0	32.4	36.5	6.8	11.3	16.1	21.2	22.9
Task 1	RoI Transformer	6.2	10.4	12.8	18.4	20.9	7.4	13.6	21.8	33.9	40.6	3.7	10.6	14.5	21.7	22.8
	ReDet	5.7	10.5	14.8	18.9	20.3	9.5	12.1	17.8	34.4	42.1	4.8	9.9	14.1	20.9	21.5
	Oriented R-CNN	7.8	12.1	14.2	18.1	20.4	10.1	14.2	22.8	33.6	39.6	5.6	10.4	15.3	21.4	22.6
	ours	**11.2**	**17.9**	**22.1**	**23.7**	**26.1**	**20.4**	**26.7**	**29.1**	**39.6**	**45.7**	**10.3**	**16.2**	**18.4**	**24.6**	**26.3**

Table 4. Detection results for each novel class under 30 shot, and mAP (%) is used as evaluation measure. Red and blue represent the best and the second best.

	Split 1						Split 2					Split 3			
Method/Class	**PL**	**SH**	**HA**	**LV**	**GTF**	**ST**	**BD**	**BC**	**TC**	**RA**	**BR**	**SV**	**HC**	**SBF**	**SP**
FR-O	33.3	9.1	10.1	9.4	30.0	39.2	41.1	19.9	48.9	42.9	10.7	22.7	27.3	28.4	21.0
Giding Vertext	26.7	8.9	8.3	14.2	26.3	30.9	34.6	16.8	47.3	44.2	10.3	22.3	9.1	29.8	20.6
R3Det	27.7	6.5	10.8	7.6	27.4	35.1	26.9	15.9	43.9	36.9	6.2	23.4	30.5	30.0	15.4
Sa2net	29.6	9.1	11.6	11.9	32.0	36.6	32.7	14.9	53.4	41.3	1.1	39.2	15.9	30.5	10.8
RoI Transformer	33.4	9.1	10.6	19.8	31.3	40.7	39.6	22.4	57.6	42.7	12.1	23.2	29.7	29.4	19.9
ReDet	35.4	9.1	11.8	17.1	29.3	34.2	31.3	28.2	71.4	45.6	15.2	19.8	27.7	29.9	15.4
Oriented R-CNN	34.2	9.5	8.3	24.0	26.1	39.2	41.4	20.3	51.5	45.4	12.1	20.3	28.5	31.3	20.5
ours	**41.7**	**9.3**	**12.3**	**33.1**	**34.2**	**41.4**	**44.4**	**28.5**	**68.6**	**46.1**	**12.4**	**28.7**	25.8	**37.2**	**27.4**

Moreover, we visualized the detection results of our method, MOCA-Net, under 30 shots, and a comparative analysis was conducted between our method and the baseline Oriented R-CNN under the same settings. As shown in Figure 6, the detection results for the novel classes are presented in the first two lines, and the third and fourth lines show the detection results of more challenging situations where both base classes and novel classes exist. Compared with Oriented R-CNN detection results, the detection results of our method have fewer missed or incorrect detection instances of objects, the orientations of the bounding boxes are more accurate, and the overlap is less, which strongly supports the advantages of our method for addressing common situations in remote sensing images,

such as multi-oriented and multi-scale objects, dense object distribution, and complex backgrounds.

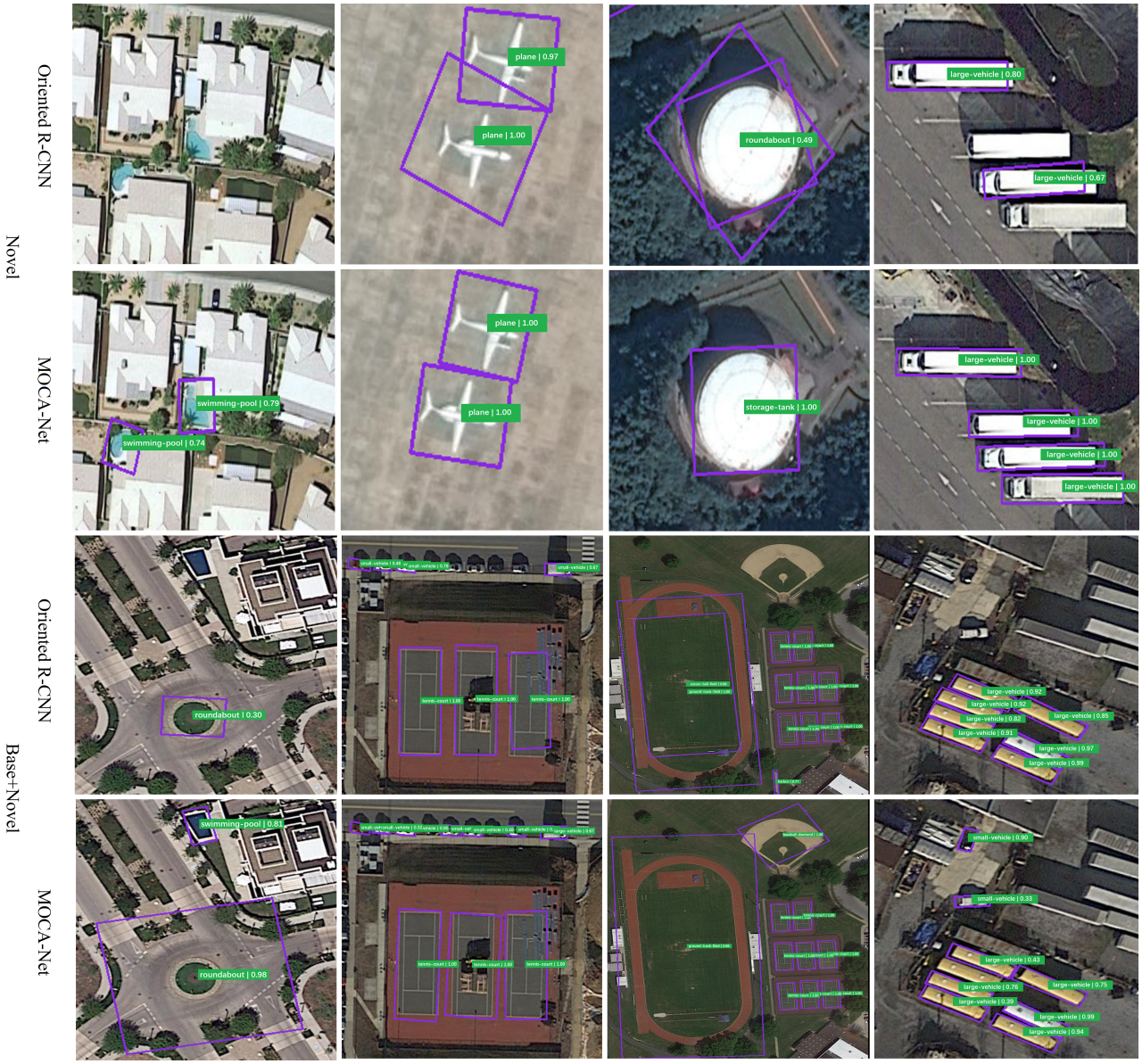

Figure 6. The visualization of detection result. Respectively, the results detected by Oriented R-CNN and MOCA-Net for the novel classes are shown in the first two rows, and the third and fourth rows show more challenging results detected by Oriented R-CNN and MOCA-Net where both novel and base classes exist.

4.5. Ablation Study

Quantitative study of sparsity in orientation and scale distribution. In Figure 2, we visualize the sparsity problem in orientation and scale distribution of few-shot oriented object detection. To further study the sparsity of orientation and scale distribution quantitatively, we constructed a small dataset with restricted orientations and scales to test our detector and compare it with the randomly selected small sample dataset we previously constructed. We selected 30 objects from the plane/small vehicle classes with scales ranging from 64^2 to 128^2 and orientations ranging from 0° to 45° for building the limited dataset. As shown in Table 5, the limitation of orientations and scales leads to a sharp decline in the detection precision of the detector, making it necessary to address the sparsity of orientation and scale distribution in few-shot oriented object detection. And our proposed method,

MOCA-Net, can significantly alleviate this problem, with less precision loss on the limited dataset compared to Oriented R-CNN.

Table 5. AP (%) in plane/small vehicle classes. Two datasets consisting of 30-shot samples are constructed on the DOTA split 1, with objects having orientations and scales that are either randomly selected or limited.

Method	Plane		Small Vehicle	
	Random	Limited	Random	Limited
Oriented R-CNN	34.2	19.3	20.3	11.8
MOCA-Net	41.7	33.6	28.7	24.5

Analysis of different settings in the multi-oriented enhancement branch. In this section, we examine the influence of different settings in the multi-oriented enhancement branch. The experiments are based on the DOTA dataset under each shot in the first base/novel class split. As shown in line 3, line 4, and line 5 of Table 6, we separately evaluated the impact of the multi-oriented enhancement branch in the RPN and the detection head (RoI) to analyze its role in the whole method. The mAP of the model with only RPN or only RoI improved by about 2% compared to the baseline in all shots, and when the two parts were combined, the mAP further improved by around 3%. And, as shown in line 2 and line 5 of Table 6, the model with both classification enhancement and regression enhancement increased the mAP by an average of 1.2% compared to the model with only classification enhancement. In the last two lines of Table 6, we explored the operation of random clipping by moderately shifting the ground truth. After adding it to the multi-oriented enhancement branch, the mAP increased by an average of 1.1%, achieving the best performance in Table 6.

Table 6. Experiment results of different settings in multi-oriented enhancement branch.

RPN		RoI			Shot				
cls	reg	cls	reg	random crop	3	5	10	20	30
					7.8	12.1	14.2	18.1	20.4
✓		✓			8.9	14.4	17.8	19.8	22.9
✓	✓				8.6	13.5	16.4	19.3	22.4
		✓	✓		8.9	13.9	16.8	19.5	22.6
✓	✓	✓	✓		9.8	15.4	19.7	21.2	23.7
✓	✓	✓	✓	✓	10.6	16.7	20.9	22.3	24.8

Contribution of different parts in context-aware module. By conducting comparative experiments, we aimed to explore the contribution of background context information and horizontal box context information in the context-aware module. The experiments were based on the DOTA dataset under each shot in the first base/novel class split. From Table 7, we can see that adding background context information or adding horizontal box context information increased the mAP by averages of 1.4% and 1.8%, respectively, and combining the two can achieve the best improvement effect, increasing the mAP by an average of 2.5%.

Table 7. Experiment results of different parts in context-aware module.

HBB	BG	Shot				
		3	5	10	20	30
		7.8	12.1	14.2	18.1	20.4
✓		8.8	13.3	16.1	19.2	21.9
	✓	8.9	13.6	16.8	19.9	22.5
✓	✓	9.3	14.2	17.4	20.8	23.4

Sensitivity and impact of expansion factor. We further explored the impact of the expansion coefficient in the extraction of background context information. The experiments

were also based on the DOTA dataset under each shot in the first base/novel class split. As shown in Table 8, when the expansion coefficient is 1.5, the model achieves the best performance. When the expansion coefficient is 1.2, the mAP decreases by an average of 1.1% compared to when the expansion coefficient is 1.5. When the expansion coefficient is 2.0, the mAP decreases by an average of 0.6% compared to when the coefficient is 1.5. Therefore, it is important to reasonably select the proportion containing the background.

Table 8. The sensitivity and impact of expansion factor in context-aware module.

r	Shot				
	3	**5**	**10**	**20**	**30**
1.2	8.2	12.7	15.3	18.8	21.2
1.5	8.9	13.6	16.8	19.9	22.5
2	8.5	13.2	15.9	19.4	21.7

Context-aware module in different methods. Because our context-aware module is plug-and-play, it can be applied to the detection heads of other multi-stage detection methods. As shown in Table 9, we add it to RoI Transformer, ReDet, and Oriented R-CNN for comparison, and it can be observed from the experimental results under each shot in the first base/novel class split that RoI Transformer, ReDet, and Oriented R-CNN show an average improvement in detection accuracy of over 2%, which proves the generality of this module among different methods.

Table 9. The effectiveness of context-aware module in different oriented object detection models.

Method	Shot				
	3	**5**	**10**	**20**	**30**
RoI Transformer	6.2	10.4	12.8	18.4	20.9
RoI Transformer w/CAM	7.9	12.2	14.3	20.5	23.1
ReDet	5.7	10.5	14.8	18.9	20.3
ReDet w/CAM	7.4	12.7	17.9	21.6	22.8
Oriented R-CNN	7.8	12.1	14.2	18.1	20.4
Oriented R-CNN w/CAM	9.3	14.2	17.4	20.8	23.4

Verification of the stability of our method. In order to obtain stable experiment results and verify the stability of our model, we repeatedly tested it three times. Table 10 shows the three experimental index values and their average performance and standard deviation. It can be seen that the gap between the results of a single experiment and the average results is mostly within 0.5%, and the standard deviation is also basically within 0.5%, which demonstrates the stability of our method, MOCA-Net.

Table 10. Stability verification experiment of our method

	Split 1					**Split 2**					**Split 3**				
Shot	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**	**3**	**5**	**10**	**20**	**30**
1#	11.4	17.9	21.8	23.8	26.4	19.6	26.1	29	40.1	45.5	9.7	16.9	18.7	24.7	26.4
2#	10.9	17.7	22.5	23.5	26.3	20.1	27.6	29.8	39.2	46.2	10.6	15.7	17.9	24.8	26.1
3#	11.3	18.1	22	23.8	25.6	20.9	26.4	28.5	39.5	45.4	10.6	16	18.6	24.3	26.4
Average	11.2	17.9	22.1	23.7	26.1	20.4	26.7	29.1	39.6	45.7	10.3	16.2	18.4	24.6	26.3
Standard deviation	0.21	0.16	0.29	0.14	0.35	0.57	0.65	0.53	0.37	0.34	0.42	0.51	0.35	0.21	0.14

5. Discussion

Our proposed method, MOCA-Net, was evaluated in numerous experiments on the DOTA datasets and compared with various general oriented object detection methods and FSD methods. The results obtained from the comparative experiments serve as evidence for the strengths of our proposed method, which will be individually analyzed and discussed in this section.

Firstly, the observation can be made from Table 2 that the performance of general oriented detection methods drops sharply when there are only few data of novel classes. For the DOTA dataset, these methods can only achieve an mAP of 20% to 40% in the 30-shot case, and the performance is significantly worse than when there are huge amounts of data. In addition, the results in Table 2 also reveal that FR-O, only adding angle θ to the last output of Faster R-CNN, performs better than Gliding Vertex, R3Det, and Sa2Net, which are well-designed for oriented object detection with large amounts of data. This situation reveals the weakness of the existing oriented object detection methods: the existing improvements proposed are mainly aimed at situations where there are plenty of annotated samples for model training, and these improvements will not significantly improve the detection accuracy when there is a lack of annotated samples. And the performance of our method surpasses all others in every case, with an mAP that was generally higher than other methods by at least 5% and even up to 10%, which demonstrates the significant advantages of introducing a multi-oriented enhancement branch and context-aware module in few-shot oriented object detection.

Secondly, from the data results presented in Table 4, the FSD method, which has excellent performance in natural images, is not significantly effective in remote sensing images. The detection accuracy of FSD methods is not significantly higher than that of the latest general oriented object detection methods such as Oriented R-CNN, while our method, MOCA-Net, performs the best in all cases, reflecting that the FSD methods for the horizontal bounding box are not completely applicable to object detection in remote sensing images, and our method is specifically designed on an oriented bounding box, making it more suitable than FSD methods. Furthermore, compared with MPSR, which only focuses on the problem of object scale, our approach increased mAP by averages of 4.7%, 11.2%, and 5.4% in split1, split2, and split3 settings, respectively. And compared with FSCE, which only focuses on the classification part, our approach increased mAP by averages of 4.9%, 9.6%, and 3.5% in split1, split2, and split3 settings, respectively. These results further demonstrate the advantages of our approach in addressing the issues of both object orientation and object scale while not only enhancing the model's regression capability but also the model's classification capability.

Thirdly, it is apparent from Table 5 that our method has exhibited outstanding performance in most categories, and for some categories, such as plane (PL) and large vehicle (LV), which have various scales and arbitrary orientations, our method has resulted in significant increases in the detection accuracy of 7.5% and 9.1%, respectively, compared to Oriented R-CNN. And it is also apparent that the detection difficulty varies across different categories in few-shot oriented object detection and that there is a significant difference in detection accuracy between different categories. For the categories with large aspect ratios, such as harbor (HA) and bridge (BR), it is still challenging for our approach to detect them. In addition, some methods have shown excellent detection performance for specific categories. For example, Sa2net, which introduces a feature alignment module, performs well on small vehicles (SV), and ReDet, which incorporates a rotation-invariant backbone and rotation-invariant ROI align, achieves outstanding results on tennis courts (TC), which is worthy of further research.

But, in few-shot oriented object detection, due to the large number of parameters and strong representation ability of CNN-based models, if the annotated samples are few, the model is easily able to remember non-universal features of the object, resulting in overfitting and reducing the model's generalization ability [12]. The results in Table 2 show that the mAPs of all methods, including our proposed method, are still less than 50% and that there is still a huge improvement range. And because of the variable scale and orientation of objects as well as the complex background in remote sensing images, few-shot object detection is more difficult in remote sensing images. It can been seen from Table 3, that the performance of FSD methods such as TFA and Meta R-CNN in remote sensing images is far inferior to their performance in natural images. In addition, for the categories with large aspect ratios, such as harbor (HA) and bridge (BR), with difficulty

in bounding box regression, it is still hard for the existing methods to detect them. And the large-scale dataset DOTA, containing 15 classes and 188,282 objects, which covers most cases in oriented object detection, makes the detection more difficult. Therefore, few-shot oriented object detection still remains a huge challenge and has great research space in the future.

Meanwhile, we conducted a comprehensive ablation study to further explore our methods. These experiments in the ablation study provide a detailed demonstration of the effects and performance of each component in the multi-oriented enhancement branch and the context-aware module.

According to the ablation experiments of the multi-oriented enhancement branch, as shown in Table 6, the models with only RPN and RoI partially enhanced all exceeded the baseline in all shots, which proved their effectiveness. When they are combined, the detection accuracy of the model is further improved, which shows that the two parts play a complementary role. And, although most FSD methods assume that the box regression capability of the model is class-agnostic and does not need to be enhanced during model transfer, it can be seen from the comparison between line 2 and line 5 of Table 6 that further improvements can be made to enhance the detection performance by incorporating regression enhancement when classification enhancement is carried out, indicating that the enhancement of the regression task is also important in few-shot oriented object detection. In addition, after adding random crop to the multi-oriented enhancement branch, the model achieved the best performance, proving that random clipping can increase the generalization of the detector and improve its detection effect.

According to the ablation experiments of the context-aware module, as shown in Table 7, we can see that adding background context information or horizontal box context information to assist the object classification can bring about a significant improvement in the detection accuracy of the detector, and that the combination of both components results in the best performance, indicating their complementary roles. And, as shown in Table 8, it can be seen that increasing the expansion coefficient does not necessarily lead to better results. When the expansion coefficient is 1.5, the module achieves the best performance. When the expansion coefficient is too small to be 1.2, the added contextual information is small, resulting in only a small increase in detection precision. When the expansion coefficient is too large to be 2.0, it will add too much contextual information, reducing the proportion of object features, and the detection precision is lower than when the coefficient is 1.5. In addition, as evidenced by the results presented in Table 9, the introduction of our context-aware module into different oriented object detection methods has led to a significant improvement in the detection accuracy after introducing our context-aware module into different oriented object detection methods, which demonstrates the importance of contextual information in few-shot oriented object detection.

For the applicability of our proposed method on real measured RS data, we tested the effectiveness of our proposed method on the public large-scale dataset DOTA (Dataset for Object Detection in Aerial Images) [20]. This dataset is the most representative public remote sensing image dataset for oriented object detection and consists of 2806 optical aerial images and contains 188,282 objects and 15 object classes, with data from various sensors and platforms such as Google Earth. Because the dataset DOTA covers most cases of oriented object detection in optical aerial images, many famous object detection algorithms in the field of remote sensing are based on this dataset, such as RoI transformer [6] and ReDet [8], and based on these methods, excellent methods have been developed for specific applications such as ship detection [18,37]. Due to this, we believe that our proposed method can be applied to scenes and inspire future research for few-shot oriented object detection in optical aerial images.

To summarize, we designed a few-shot oriented object detection method, MOCA-Net, based on a multi-oriented enhancement branch and a context-aware module in this study, and our method has been proven effective through numerous experiments and an ablation study. Although a remarkable level of improvement is brought about through our method,

as shown in the experiment results, the detection accuracy of our method is still relatively low, with an mAP of less than 40% in most cases. And the multi-oriented enhancement branch cannot be fully plug-and-play like the context-aware module, which makes it unable to be directly added to other methods such as RoI transformer and ReDet. In the future, we will make the multi-oriented enhancement branch easier to add to other methods and improve our approach by studying the principles of other methods, such as Sa2net and ReDet, refining the network architecture, and optimizing the training strategy to further enhance the detection performance on various categories, with a specific focus on categories characterized by a large aspect ratio.

6. Conclusions

In this study, we present MOCA-Net, a method specifically designed for few-shot oriented object detection, and is the first attempt at this in this field. To address the issue of sparse orientation and scale distributions, MOCA-Net generates multi-oriented and multi-scale positive samples and avoids introducing too many improper negative samples. It enhances the classification and regression capabilities for various orientations and scales of objects in the RPN and the detection head, forming a multi-oriented enhancement branch. Simultaneously, with the aim of improving the detector's classification capability, the method introduces the context-aware module in the detection head, and through the integration of contextual information into the RoI features, the method alleviates the classification confusion problem when there is little data on novel classes for training, leading to enhanced detection accuracy. In comparative experiments on the public oriented object detection dataset DOTA, the detection performance of our method surpasses that of both general oriented object detection methods and existing few-shot object methods, which proves its effectiveness. Few-shot oriented object detection in remote sensing images poses significant challenges, and our method still has limitations, so our next study aims to delve deeper into this field to achieve enhanced performance by studying the principles of other methods, refining the network architecture, optimizing the training strategy, and so on.

Author Contributions: Conceptualization, H.S.; methodology, H.S.; software, H.S.; validation, S.L.; formal analysis, H.S.; investigation, Y.Y.; data curation, H.S.; writing—original draft, H.S.; writing—review & editing, S.L.; project administration, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62101060).

Data Availability Statement: The source code generated and used for this study is publicly available at https://github.com/xuehua-piaopiao/MOCA-Net.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Liu, W.; Anguelov, D.; Erhan, D.; et al. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
3. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8232–8241.
6. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.

7. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
8. Han, J.; Ding, J.; Xue, N.; Xia, G.S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
9. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
10. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3630–3638.
11. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090.
12. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8420–8429.
13. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9577–9586.
14. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5192–5201.
15. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
16. Ang, X.; Huang, T.; Gonzalez, J.; Darrell, T.; Yu, F. Frustratingly Simple Few-Shot Object Detection. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 9919–9928.
17. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 7352–7362.
18. You, Y.; Ran, B.; Meng, G.; Li, Z.; Liu, F.; Li, Z. OPD-Net: Prow Detection Based on Feature Enhancement and Improved Regression Model in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 6121–6137. [CrossRef]
19. Fu, K.; Chang, Z.; Zhang, Y.; Sun, X. Pointbased estimator for arbitrary-oriented object detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4370–4387. [CrossRef]
20. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–21 June 2018; pp. 3974–3983.
21. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 142–149. [CrossRef]
22. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 3520–3529.
23. Li, J.; Wei, Y.; Liang, X.; Dong, J.; Xu, T.; Feng, J.; Yan, S. Attentive contexts for object detection. *IEEE Trans. Multimed.* **2018**, *19*, 944–954. [CrossRef]
24. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 456–472.
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
26. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
27. Liao, M.; Shi, B.; Bai, X. Textboxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
28. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]
29. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4013–4022.
30. Han, G.; Huang, S.; Ma, J.; He, Y.; Chang, S.F. Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 780–789.
31. Han, G.; Ma, J.; Huang, S.; Chen, L.; Chang, S.F. Few-shot object detection with fully cross-transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5321–5330.
32. Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; Savvides, M. Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8782–8791.

33. Cai, J.; Zhang, Y.; Guo, J.; Zhao, X.; Lv, J.; Hu, Y. St-pn: A spatial transformed prototypical network for few-shot sar image classification. *Remote Sens.* **2021**, *14*, 2019. [CrossRef]
34. Rostami, M.; Kolouri, S.; Eaton, E.; Kim, K. 2019. Deep transfer learning for few-shot SAR image classification. *Remote Sens.* **2021**, *11*, 1374. [CrossRef]
35. Tai, Y.; Tan, Y.; Xiong, S.; Sun, Z.; Tian, J. Few-shot transfer learning for sar image classification without extra sar samples. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2240–2253. [CrossRef]
36. Wang, L.; Bai, X.; Zhou, F. Few-shot SAR ATR based on conv-BiLSTM prototypical networks. In Proceedings of the 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Xiamen, China, 26–29 November 2019; pp. 1–5.
37. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087. [CrossRef]
38. Zhang, H.; Zhang, X.; Meng, G.; Guo, C.; Jiang, Z. Few-Shot Multi-Class Ship Detection in Remote Sensing Images Using Attention Feature Map and Multi-Relation Detector. *Remote Sens.* **2022**, *14*, 2790. [CrossRef]
39. Zhang, S.; Song, F.; Liu, X.; Hao, X.; Liu, Y.; Lei, T.; Jiang, P. Text Semantic Fusion Relation Graph Reasoning for Few-Shot Object Detection on Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1187. [CrossRef]
40. Chen, S.; Zhang, J.; Zhan, R.; Zhu, R.; Wang, W. Few Shot Object Detection for SAR Images via Feature Enhancement and Dynamic Relationship Modeling. *Remote Sens.* **2022**, *14*, 3669. [CrossRef]
41. Su, H.; You, Y.; Meng, G. Multi-Scale Context-Aware R-Cnn for Few-Shot Object Detection in Remote Sensing Images. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1908–1911.
42. Gidaris, S.; Komodakis, N. Object detection via a multi-region and semantic segmentation-aware CNN model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1134–1142.
43. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
44. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. MMrotate: A rotated object detection benchmark using pytorch. In Proceedings of the ACM International Conference on Multimedia, New York, NY, USA, 10–14 October 2022; pp. 7331–7334.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 3163–3171.
47. Xiao, Y.; Marlet, R. Few-shot object detection and viewpoint estimation for objects in the wild. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 192–210.